

We Claim:

1. A method for determining a degree of similarity between documents, the method comprising the steps of:
 - 5 storing, for at least two documents, labeled tree representations of respective documents;
 - 10 storing, for at least two documents, path representations relating to paths that occur in the documents from root nodes to leaf nodes in the labeled tree representations of the respective documents; and
 - 15 calculating a measure of similarity between two of the documents based upon the frequency of occurrence of similar paths specified by the path representations.
2. The method as claimed in claim 1, wherein the tree representation is a Document Model Object representation.
- 20 3. The method as claimed in claim 1, further comprising the step of generating a path representation for a path of a document as a sequence of labels representative from a root node to a leaf node in the labeled tree representation of the document.
- 25 4. The method as claimed in claim 1, further comprising the step of storing, as path representations, sets of sequenced labels representative of distinct paths in a labeled tree representation of a corresponding document.
- 30 5. The method as claimed in claim 4, further comprising the step of storing a path dictionary ($Dict_{paths} = \{p_1, p_2, \dots, p_N\}$) of distinct paths collated from a tree representation for a document.

6. The method as claimed in claim 5, further comprising the step of eliminating selected paths from the path dictionary ($Dict_{paths}$).
7. The method as claimed in claim 6, wherein paths that occur highly frequently or highly infrequently are eliminated from the path dictionary ($Dict_{paths}$).
8. The method as claimed in claim 7, further comprising the step of computing the frequency of occurrence ($f_j(p_i)$) of a path (p_i) in a document (d_j).
9. The method as claimed in claim 8, further comprising the step of computing the maximum number of instances ($f_{max} = \max_{ij} f_j(p_i)$) in which a path (p_i) in the document (d_j) occurs.
10. The method as claimed in claim 9, further comprising the step of storing a representation of the document (d_j) as a N -dimensional vector ($[d_{j1}, d_{j2}, \dots, d_{jN}]$, where $d_{jk} = f_j(p_k)/f_{max}$, $1 \leq k \leq N$) of relative frequencies of occurrence ($f_j(p_k)$) of paths (p_k) in the document (d_j).
11. The method as claimed in claim 8, further comprising the step of computing the minimum number of instances ($f_{min} = \min_{ij} f_j(p_i)$) in which a path (p_i) in the document (d_j) occurs.
12. The method as claimed in claim 10, further comprising the step of computing the similarity between a pair of documents (d_i, d_l) as a function ($sim(d_i, d_l)$) of metrics relating the number of paths common to the respective documents (d_i, d_l).
13. The method as claimed in claim 12, wherein the function for computing the similarity between a pair of documents (d_i, d_l)

$$(sim(d_i, d_l) = sim(d_i, d_l) = \frac{\sum_{k=1}^N \min(d_{ik}, d_{lk})}{\sum_{k=1}^N \max(d_{ik}, d_{lk})})$$

5 is the quotient of a numerator, defined as the sum for all paths ($k = 1 \dots N$) of the minimum number of instances ($\min(d_{ik}, d_{lk})$) in which paths occur in the respective documents (d_i, d_l), and a denominator, defined as the sum for all paths ($k = 1 \dots N$) of the maximum number of instances ($\min(d_{ik}, d_{lk})$) in which paths occur in the respective documents (d_i, d_l).

10 14. The method as claimed in claim 1, wherein the tree representation of a document includes a positional index, which represents, for a node (n), the number of previous sibling nodes with the same label as that of node (n).

15 15. The method as claimed in claim 14, further comprising the step of storing as a path representation a set that defines positional information of sibling nodes under a parent node.

16. The method as claimed in claim 15, further comprising the step of storing precise path representations that precisely define a document structure, and generalised path representations that partially generalise structural aspects of precise path representations of a document.

20 17. The method as claimed in claim 16, wherein the step of calculating the measure of similarity involves determining a total number of precise path representations of one document that are either shared by the other document, or are a subsumed subset of at least one of the generalised path representations of the other document.

25 18. The method as claimed in claim 17, further comprising the step of normalising the measure of similarity by a term that represents the number of unique path representations shared by the two documents.

30 19. The method as claimed in claim 18, wherein the number of unique path representations is calculated by adding the number of path representations for

each document, and subtracting from this total the number path representations shared by the two documents.

20. The method as claimed in claim 14, further comprising the step of storing as a
5 path representation a sequence of terms separated by a delimiting symbol, in which each term is represented by a label and a parenthesised predicate that specifies the positional index of the term either specifically or generally.

21. Computer software, recorded on a medium, for determining a degree of
10 similarity between documents, the computer software comprising:

software code means for storing, for at least two documents, labeled tree representations of respective documents;

15 software code means for storing, for at least two documents, path representations relating to paths that occur in the documents from root nodes to leaf nodes in the labeled tree representations of the respective documents; and

software code means for calculating a measure of similarity between
20 two of the documents based upon the frequency of occurrence of similar paths specified by the path representations.

22. A computer system for determining a degree of similarity between documents,
the computer system comprising:

25 means for storing, for at least two documents, labeled tree representations of respective documents;

means for storing, for at least two documents, path representations
30 relating to paths that occur in the documents from root nodes to leaf nodes in the labeled tree representations of the respective documents; and

means for calculating a measure of similarity between two of the documents based upon the frequency of occurrence of similar paths specified by the path representations.